CSCI 567: Machine Learning

Spring 2020

Class Project: Predict Future Sales

1 Introduction

For this project, you will be competing on a Kaggle competition to learn to predict on a challenging time-series dataset consisting of daily sales data, originally provided by a Russian software firm. The task is to predict total sales for every product and store for the upcoming month. Please see the Kaggle link: https://www.kaggle.com/c/competitive-data-science-predict-future-sales/overview for more details and to download the dataset.

2 Team formation

Each team should consist of maximum 3 members. A team size larger than this is only permissible under special circumstances and requires the instructor's explicit approval. Please form a team of upto **3** students and fill out your team details in this sign-up google doc by **Feb 25, 11:59pm PT**. Only one member of the team needs to fill up the google doc and sign up your full team for the project.

- A team can include both in-class and DEN students. You can use the online project forum on DEN to look for team members if needed. However, teams can only include students registered for CSCI-567 Spring 2020 at USC. No outside members are allowed to be a part of the team.
- Absolutely no changes to the team composition (e.g. adding members, swapping members, removing members, team mergers etc.) will be entertained after the sign-up deadline so please choose your team wisely. Even if a team member withdraws from the class or is not able to complete the class or work on the project due to any reason, new teammates won't be permitted.
- The name of your team in the form should be the same as that on the Kaggle leaderboard and must begin with `CS567_`, e.g. `CS567_WeLoveML`.
- All members of the same team will receive the same total grade for the project.
- All solutions for a team must be posted from their team account thereby adhering to Kaggle's limit of max **5** solutions per day. Creating additional Kaggle accounts to post more solutions per day is forbidden during this competition. Any student/team caught violating this rule will be immediately disqualified from the competition and will potentially face more severe penalties.

3 Grading

Your grade for the project is based on three components: (1) your team's relative score and ranking on the leaderboard (**60%**), (2) the final project presentation (**20%**), and (3) the project report (**20%**).

- Note that the leaderboard shows the ranking of all teams, not only those in this class. We will take the relative ranking among all teams into consideration. A better ranking will always lead to a better grade.

- In your presentation, you should briefly summarize the dataset exploration, the learning algorithms you tried and the results obtained along with a thoughtful comparison of your approaches.
- In your report, you should cover the details of your solutions, including the general ideas, the data processing and cleaning, the learning algorithms and models your team tried, the results you obtained and any other insightful thoughts during the competition.
- Each team's final report needs to be written in LaTeX in the **NeurIPS format** (6 pages maximum, including all references; this page limit is strict). The style files for the NeurIPS LaTeX template can be found here: https://neurips.cc/Conferences/2019/PaperInformation/StyleFiles. Not following the template will lower your grade. The report needs to be submitted online on DEN (submission link will be released at a later date).
- We will record the scores and ranking on the leaderboard on **Apr 27, 11:59pm PT** for grading.

4 Deadlines

- Project team sign-up deadline: **Feb 25, 11:59pm PT**
- Leaderboard recorded for grading: **Apr 27, 11:59pm PT**
- Final presentations: In-class on **Apr 29 and May 1, 10am-11:50am PT**
- Report submission deadline: **May 10, 11:59pm PT**

5 Additional information

- No computing credits will be provided for this project. The dataset size is less than 100 MB and can be easily handled on a personal laptop without requiring expensive GPUs. However, you are free to setup computing accounts on AWS, Google Cloud, Microsoft Azure etc. if you need to train your models online.
- We recommend that you start early and post solutions on a regular basis, since your team can only upload maximum 5 solutions per day. Trying to work on the project just before the deadline may not work out well due to this restriction.
- `scikit-learn` can be a good starting point if you want to try out reliable implementations of existing ML algorithms. There are also many notebooks exploring the dataset available on Kaggle. Several clarifications can be obtained on the Kaggle discussion forums for the competition.